Применение методов классификации для анализа визитных карточек в мобильном телефоне

Т.В. Гордеева, tangord@mail.ru, студентка (Санкт-Петербургский государственный электротехнический университет «ЛЭТИ», ул. Профессора Попова, 5, г. Санкт-Петербург, 197376, Россия)

В статье рассмотрены подходы к решению задачи классификации, применяемые для анализа визитных карточек в мобильном телефоне. Выявлены достоинства и недостатки существующих приложений, специализирующихся на сканировании визиток. Предложена математическая модель выбранного подхода. Рассмотрен вариант архитектуры предлагаемого решения. Представлены результаты классификации на нескольких примерах. Сделаны выводы о работоспособности подхода и дальнейших улучшениях решения.

Ключевые слова: Business Card Reader, OCR, многоклассовая классификация, Наивный байесовский классификатор, метод опорных векторов, математическая модель.

Визитная карточка – лицо делового человека. В современном мире она является инструментом, который отражает информацию о человеке или организации, такую как имя, фамилия, email, телефон, название организации и т.д. Зачастую количество визиток прибавляется с каждым днём, поэтому становится всё сложнее их хранить и искать важные контакты.

Данную проблему помогает решать Business Card Reader (BCR), основанный на задаче распознавания визиток. Задача BCR – дать людям возможность извлечь информацию из визитных карточек, не прибегая к ручной установке адресов, телефонов, имен.

Процесс распознавания визиток проходит в несколько этапов:

– Извлечение текста из изображений при помощи камеры мобильного телефона и OCR (Optical Character Recognition);

– Разделение полученного текста на категории.

В настоящее время задача сканирования визиток реализована в платном приложении Business Card Reader Free от ABBYY [1], которое предоставляет бесплатно распознавание десяти визиток. BCR Free предлагает пользователям переносить данные визиток на 25-ти языках и имеет обширную словарную базу, что увеличивает точность распознавания. При этом в приложении нет возможности распознать заново уже имеющуюся визитку при ее редактировании.

Платные приложения Business Card Reader Pro [2] от SHAPE, ScanBizCards [3] от ScanBiz Mobile Solutions L.P., WorldCard Mobile [4] от Penpower Technology Ltd. доступны для устройств с iOS.

В основном, многие существующие решения являются платными и и поддерживаются не всеми платформами, а в бесплатных версиях не всегда включен весь необходимый функционал. Учитывая все вышеизложенные факты, можно сделать вывод, что разработка нового приложения, компенсирующего все недостатки имеющихся решений, является актуальной задачей.

Для извлечения текста из визитки можно воспользоваться любым OCR, например Tesseract [5]. При этом точность полученного результата снижается, если визитка напечатана на сложном фоне, поэтому следует применять алгоритмы первоначальной обработки изображения для изоляции текста [6,7].

Следующим шагом является классификация полученного текста. При этом нет возможности перечислить все возможные варианты в словаре, так как существует большое количество имен, фамилий, организаций и т.д. Поэтому необходимо выполнять автоматический разбор категорий.

Рассматриваемые подходы к решению задачи классификации используют обучение с учителем [8]: на обучающем наборе слов вычисляются статистические параметры модели, а затем данные параметры используются для предсказания класса неизвестных слов.

Текст визитной карточки может содержать большое количество категорий, потому задача сводится к многоклассовой классификации, которая имеет следующую формальную постановку:

X – пространство объектов;

Y – множество ответов;

y: X →Y – неизвестная целевая зависимость, значения которой известны только на объектах конечной обучающей выборки

Требуется построить алгоритм a: X →Y, который способен классифицировать произвольный объект на всём пространстве X.

Для решения требуется свести задачу к бинарной классификации. Рассмотрим следующие стратегии:

– «Один-ко-многим»

Идея стратегии состоит в построении М классификаторов, которые отделяют каждый класс от остальных. Получим М задач бинарной классификации. Вычисляем оценки принадлежности каждому классу .

Тогда решающее правило принимает вид:



Данный алгоритм строит линейное число классификаторов, каждый из которых обучается на полной выборке.

– «Многие-ко-многим»

Идея стратегии состоит в построении классификаторов для каждой пар классов. Получим задач бинарной классификации. Вычисляем оценки принадлежности каждому классу

Тогда решающее правило принимает вид:



Данный алгоритм строит квадратичное число классификаторов, при этом каждый из классификаторов обучается на небольшой подвыборке.

Рассмотрим некоторые алгоритмы, применяемые в задаче классификации.

– Наивный байесовский классификатор.

В основе классификатора лежит теорема Баейса.

, где

– вероятность, что объект x принадлежит классу k;

– вероятность встретить объект x среди всех объектов класса k;

 – безусловная вероятность встретить объект класса k в корпусе объектов;

– безусловная вероятность документа объекта x в корпусе объектов.

Далее требуется рассчитать вероятность всех классов и выбрать класс с наибольшей вероятностью, т.е.



Предполагаем, что признаки объекта х зависят от класса k и не зависят друг от друга:



Теперь формула классификатора принимает вид:



– Метод опорных векторов.

Рассмотрим задачу бинарной классификации.

Пусть объекты представлены векторами .  Тогда классификатор принимает вид:

, где

– признаки объекта x.

, – параметры алгоритма.

описывает разделяющую классы в пространстве гиперплоскость.

Данный метод допускает многоклассовую классификацию с применением стратегии «Один-ко-многим».

Для выбора наиболее подходящего алгоритма для анализа визитных карточек в мобильном телефоне следует учитывать следующие критерии:

– время обучения;

– качество классификации.

Алгоритмы были протестированы на размеченном корпусе фамилий, имен и отчеств на русском языке из 4500 слов. Набор признаков объектов был определен как первая и последние три буквы.

Процесс обучения построен следующим образом: 70% примеров из обучающей выборки использованы для вычисления параметров модели, 30% - для оценки качества классификатора. Также предварительно были подобраны наилучшие параметры алгоритмов (для метода опорных векторов) с помощью кросс-валидации. Наилучшие показатели классификации были выявлены при использовании Наивного Байесовского алгоритма со стратегией «один-ко-многим» (табл.1).

|  |  |  |  |
| --- | --- | --- | --- |
| Алгоритм/ стратегия | Время обучения | F-мера | |
| micro | macro |
| Наивный байесовский классификатор/ один-ко-многим | 0.037 c | 88,96% | 88,6% |

*Таблица. 1. Наилучшие показатели классификатора*

Опишем решение в виде следующей математической модели: M=(X, Y, Xl, G, A, K, W, F, P, H), где X – множество классифицируемых слов; Y – множество допустимых ответов; Xl – множество объектов обучающей выборки; G = y\*: X →Y – неизвестная целевая зависимость, значения которой известны только на объектах конечной обучающей выборки ; A: X →Y – алгоритм приближения неизвестной целевой зависимости на всем множестве X; K = {k} – множество классов; W – множество признаков каждого класса; F: X → W – функция извлечения признаков W из классифицируемых объектов; P – множество вероятностей принадлежности классифицируемого слова каждому классу; H – множество классификаторов А.

Рассмотрим работу математической модели на примере. Пусть имеется множество классифицируемых слов X, состоящих из имен фамилий и отчеств на русском языке, тогда множество K имеет вид K = {Фамилия, Имя, Отчество}. Пусть множество Y имеет вид Y = {0,1,2}, 0 – если x∈X является именем, 1 – фамилией, 2 – отчеством.

В качестве множества признаков W выделим первый и последний три символа каждого слова.

На первом этапе обучим 3 классификатора на конечной размеченной обучающей выборке , каждый из которых отделяет один из классов от остальных.

Далее классифицируем следующие объекты {‘Иванов’, ‘Иван’, ‘Иванович’}. Тогда множество признаков принимает вид W = {‘И н о в’, ‘И в а н’, ‘И в и ч’}. Далее, по множеству признаков W вычислим вероятности принадлежности (P ) слов каждому классу. Для каждого слова выберем класс с максимальной вероятностью. Тогда множество ответов принимает вид {0, 1, 2}.

Учитывая предложенную математическую модель и требования к решению разработана следующая архитектура системы распознавания визитных карточек (рис.1).

– Уровень представления

На данном уровне выполняется считывание изображение, выполняется запрос на передачу данных серверу для распознавания и редактирование классифицированных данных, выполняется запрос на сохранение результатов распознавания визитки. Также осуществляется просмотр и редактирование уже отсканированных и сохраненных визиток.

– Модуль проверки данных

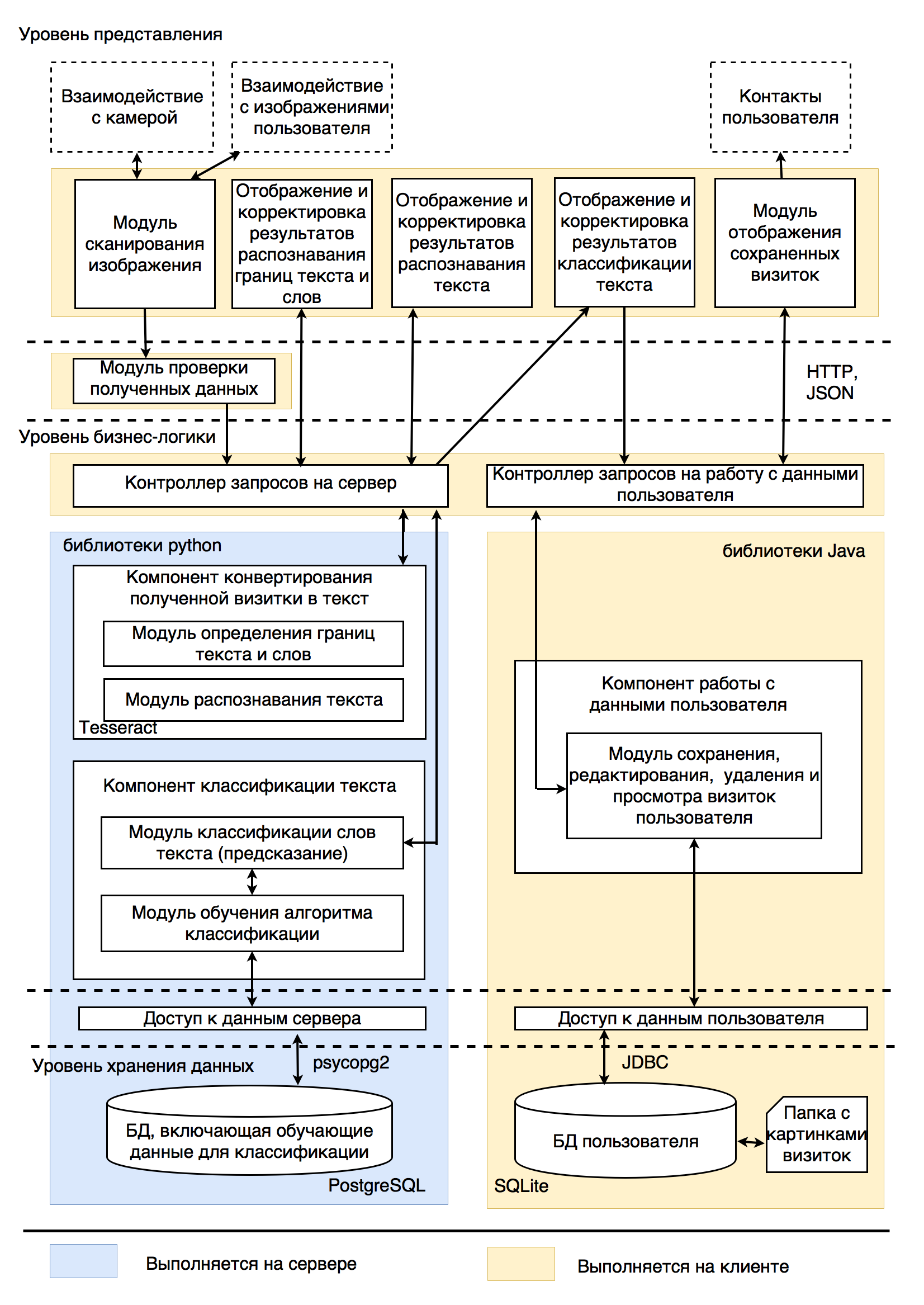
Осуществляет проверку полученного от пользователя изображения.

– Уровень бизнес-логики

На данном уровне происходят операции по распознанию текста на визитках, операции по формированию ответа пользователю на запросы  классификации, сохранения, изменения и удаления данных. Также на данном уровне формируются запросы в БД.

– Уровень хранения данных

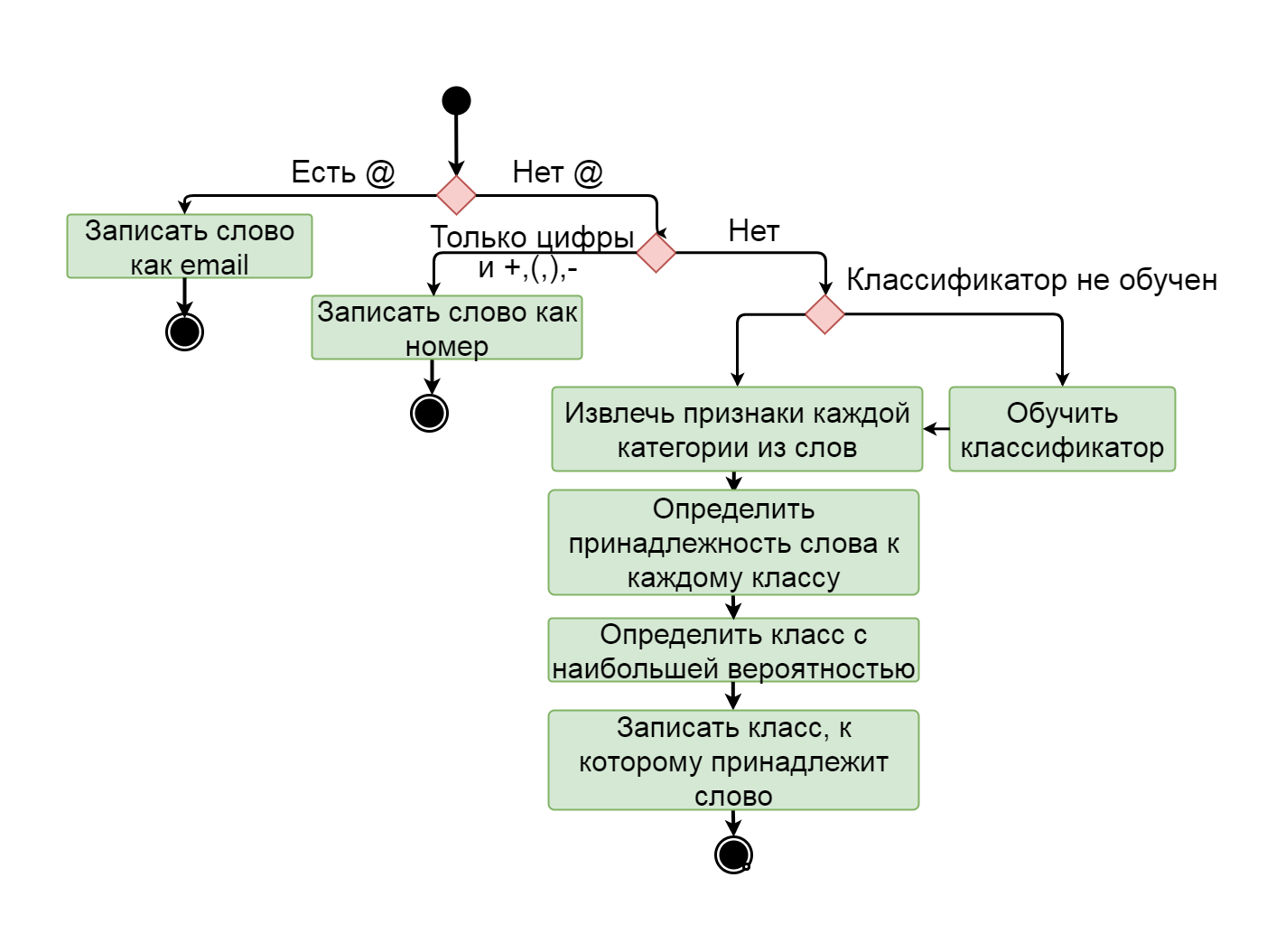
Содержит базы данных для классификации (имена, фамилии, должности и т.д.) и хранения данных пользователя (сохраненных визиток).



*Рис. 1. Архитектура решения*

Для поиска телефонов и email предлагается использовать алгоритмы на основе регулярных выражений.

Схема работы компонента классификации представлена на рисунке 2.



*Рис. 2. Схема работы компонента классификации текста.*

Обучим и протестируем предложенный алгоритм на размеченном корпусе, который состоит из имен, фамилий и отчеств в разном количественном соотношении. Результаты экспериментов представлены в таблице 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Количество слов | Соотношение классов  (имя/фамилия/отчество) | Время обучения | F-мера | |
| micro | macro |
| 4500 | 1500/1500/1500 | 0.033c | 89.6% | 89.5% |
| 7500 | 3000/3000/1500 | 0.059c | 86.1% | 88.1% |
| 15000 | 7500/6000/1500 | 0.156c | 84.8% | 87.9% |

*Таблица. 2. Результаты экспериментов*

Количество объектов обучающей выборки, принадлежащих определенному классу, влияет на точность классификации. Так как при использовании стратегии «один-ко-многим» может возникнуть проблема с несбалансированными выборками.

Предложенный подход при анализе визитных карточек позволяет проводить классификацию текста, содержащего фамилии, имена и отчества с точностью от 84%. Данный показатель может быть увеличен за счет поиска дополнительных признаков классов. Также предстоит решить задачу выбора признаков для других классов, которые могут содержаться в тексте визитки. Например, «Организация», «Должность», «Адрес» и т.д.

Литература

1. Card Reader Free. ABBYY. URL: http://www.abbyybcr.com/en/

2. Card Reader Pro. SHAPE. URL: http://www.shape.ag/en/products/details.php?product=bcr

3. ScanBizCards. ScanBiz Mobile Solutions L.P. URL: https://itunes.apple.com/us/app/scanbizcards/id335047649?mt=8

4. WorldCard Mobile. Penpower Technology Ltd. URL: http://www.penpowerinc.com/product.asp?sn=392

5. Tesseract OCR. URL: https://opensource.google.com/projects/tesseract

6. P. Nagabhushan, S. Nirmala. Text Extraction in Complex Color Document Images for Enhanced Readability. Department of Studies in Computer Science, University of Mysore, Mysore, India, 2010.

7. A. F. Mollah , S. Basu , M. Nasipuri\* , D. K. Basu. Text/Graphics Separation for Business Card Images for Mobile Devices. School of Mobile Computing and Communication, Jadavpur University, Kolkata, India.

8. Обучение с учителем. URL: http://www.machinelearning.ru/wiki/index.php?title=Обучение \_с\_ учителем

9. Байесовский классификатор. URL: http://www.machinelearning.ru/wiki/index.php?title= Байесовский\_классификатор

10. К. В. Воронцов. Лекции по методу опорных векторов.

11. Сегаран. Т. Программируем коллективный разум. // Пер. с англ. – СПб: Символ-Плюс, 2008. – 368 с.